FACULTY WORKING PAPER 92-0155

Predicting Loan Risk with an Inductive Learning Approach

James A. Gentry
Department of Finance
University of Illinois

Michael J. Shaw
Department of Business Administration
University of Illinois

Antoinette C. Tessmer
Department of Finance
University of Illinois

David T. Whitford
Department of Finance
University of Illinois

Bureau of Economic and Business Research
College of Commerce and Business Administration
University of Illinois at Urbana-Champaign

BEBR

FACULTY WORKING PAPER NO. 92-0155

College of Commerce and Business Administration

University of Illinois at Urbana-Champaign

August 1992

Predicting Loan Risk with an Inductive Learning Approach

James A Gentry
Department of Finance

Michael J. Shaw
Department of Business Administration

Antoinette C. Tessmer
Department of Finance

David T. Whitford
Department of Finance

PREDICTING LOAN RISK WITH AN INDUCTIVE LEARNING APPROACH

by

James A. Gentry
IBE Distinguished Professor of Finance
University of Illinois, Urbana-Champaign

Michael J. Shaw
Associate Professor of Business Administration
University of Illinois, Urbana-Champaign

Antoinette C. Tessmer
Visiting Assistant Professor of Finance
University of Illinois, Urbana-Champaign

and

David T. Whitford
Associate Professor of Finance
University of Illinois, Urbana-Champaign

ABSTRACT

This research determines loan risk ratings by using cash flow and qualitative information in an inductive learning system. The objective is to generate loan ratings with an inductive learning system that match the loan ratings assigned by the staff of a large regional bank. When multiple trees were developed, it became apparent that the structure of the original trees were modestly unstable, as were the positions of the attributes on the trees. A dynamic updating procedure was used to stabilize the structure of the tree as well as improve its predictive accuracy. The simplified structure of the resulting global decision tree greatly enhances the insights of the credit analysts in interpreting the attributes that underlie the loan risk ratings. The attributes selected by the global decision tree were quite similar to a hypothesized hierarchy of the cash flow components. The inductive learning system provides unique and subtle insights into the loan risk classification system, plus making it possible to have a rich and deep interpretation of the components associated with the loan risk ratings. In summary, the results indicate that inductive learning is a valuable tool for analyzing credit risk.

PREDICTING LOAN RISK WITH AN INDUCTIVE LEARNING APPROACH[1]

I. INTRODUCTION

Commercial banks develop loan risk rating systems to evaluate loan applicants and to monitor the performance of current loan customers. In general loan risk ratings are similar to the ratings assigned to bonds by credit rating agencies. A modern loan rating system is designed to have easy access to large data bases that contain both accounting and nonaccounting information. These state of the art rating systems provide the information to test various credit risk hypotheses and to discover significant relationships in the data. Also an accurate credit rating system may provide a bank with a competitive advantage in its marketplace. When a bank staff assigns a risk rating class to a loan, it signals its assessment of a company's financial health. The rating process integrates both objective and subjective information, and it provides a focal point for the bank's top management, the loan review committee, and the lending staff.

Building a loan risk rating model is a difficult and complex task. One of the reasons for this difficulty is the availability of vast quantities of accounting and nonaccounting information for analysis. Another reason is that the type of model used can affect the predictive accuracy in classifying loans correctly. We have learned through experience that banks create their own loan risk rating models. The banks may use a statistical approach, such as cluster analysis or multiple regression analysis, in conjunction with accounting and nonaccounting information to create a risk rating equation, or they may use only qualitative information, Snyder [1990]. We have found that bank management prefers a model that provides consistent and relatively accurate ratings which reflect the financial health of their existing customers and future loan prospects. Additionally, lenders prefer a model that generates unique insights into the subtle differences and nuances that exists among risky firms.

In the 1970s and early 1980s a few studies were designed to determine whether accounting information combined with selected qualitative variables could accurately predict the risk ratings that were assigned by the bank staff. This approach is called a loan risk classification study. The two leading loan risk classification studies were developed by Dietrich and Kaplan (DK) [1979] and Marais, Patell, and Wolfson (MPW) [1984]. Both studies used accounting information and qualitative data in a polytomous probit model to predict loan risk ratings assigned by a bank staff. In recent years there have been no new published studies that focused on predicting loan risk ratings.

Since the last study by MPW in 1984 two developments have added fresh insight to the classification of loan risk. First, the use of inductive learning systems provide a valuable analytical approach, Duchessi, Shawky and Seagle [1988], Shaw and Gentry [1988] and Srinivasan and Kim [1988]. Inductive learning is based on an information theory concept called "entropy." Second, relative cash flow components have been used to predict bankruptcy and bond ratings, Aziz and Lawson [1989], Casey and Bartczak [1984, 1985], Dambolena and Shulman [1088], Gentry, Newbold, and Whitford (GNW) [1985, 1989], Gombola, Haskins, Ketz and Williams [1987], Largay and Stickney [1980], and Neill, Schaefer, Bahnson and Bradbury [1991]. The cash flow information provides insights and interpretations of bankruptcy and bond ratings that are unique vis-a-vis financial ratios, GNW [1990]. This paper will use cash flow information in an inductive learning system, as well as a polytomous probit model, to predict loan risk ratings. In classifying loan risk, an inductive learning system adds value by creating a decision tree structure of the credit rating information. The hierarchical structure of the decision tree provides a tool to develop a deep interpretation of the interrelationships that exist among the components used to generate the loan risk rating.

The paper reviews the loan risk classification literature in Section II, which provides a foundation for the model building segment. Section III briefly reviews the concepts of relative cash flow components and surplus/deficit cash flow. It presents a hierarchy for using relative cash flow components to interpret a firm's financial health. Section IV develops the underlying theory of inductive learning and interprets the value that it adds in classifying loan risk. The data acquisition for the empirical analysis is found in Section V. The interpretations of the findings generated by the inductive learning system and by the probit model are in the sixth and seventh sections, respectively. A summary is presented in Section VIII.

II. LITERATURE REVIEW

Orgler [1970] developed a multiple regression model for classifying loan risk. The model used one financial ratio, net working capital/total current assets, and five dummy variables to predict whether bank examiner ratings of a loan are good, bad or marginal.[2] The Orgler model correctly predicted the ratings of 95 percent (56/59) of the good loans and of 91.1 percent (123/135) of the marginal loans. However, the model's classification accuracy was 56.6 percent (60/106) for bad loans.[3]

An objective of the two recent loan risk classification studies by Dietrich and Kaplan (DK) [1979] and Marais, Patell and Wolfson (MPW) [1984] is to develop statistical models for classifying loan risk that are based on accounting information. Both studies developed polytomous probit models which generated conditional probabilities for determining the risk rating of each loan.

The DK analysis was based on 140 companies whose financial data were on COMPUSTAT. Of the 140 companies used in determining the parameters, approximately 78 percent (109/140) were classified by the bank as being current, Category I, which DK indicate is normal, acceptable banking risk. They found three variables--debt/equity, fixed charge coverage and number of consecutive years of sales decline--classified 85 percent of all loans correctly. However, they found the loans not rated current by the bank were correctly classified less than 60 percent of the time, while the classification accuracy of the Category I loans was 93 percent. A validation test provided similar test results.

The study by MPW was based on financial data from 205 public companies and 716 private companies. They started with 20 financial variables and six nonfinancial variables for the public firms. Although approximately 93 percent of all loans were classified correctly, 90 percent of the total sample were initially rated as Category I loans by the bank staff. The results showed the misclassifications as being relatively high for loans rated other than Category I.

An empirical research project by von Stein and Ziegler [1984] focused on the prognosis and surveillance of corporate credit risks. The authors used both quantitative and qualitative measures. They presented a three-part approach that incorporated an early warning system, an evaluation of a bank-accounts information system and a system to assess the management.

A review of the literature on loan risk classification highlights several subtle dimensions. First, because public data sources are readily available, numerous studies focus on the prediction of bankruptcy and bond ratings. However, in contrast, only a few loan risk classification studies have been conducted because loan information and data are not in the public domain. Second, the information used in prior studies was from companies whose shares were either listed on a stock exchange or were private companies rated Category I. In the two studies by DK and MPW only a small percentage of the companies were rated as high risk. As expected, the accuracy of these studies in predicting the rating of low risk type loans is quite high, but these models were only modestly successful in predicting the ratings of higher risk loans. Third, prior studies commented on the need to use both quantitative and qualitative information in the prediction process. Finally, the financial variables used were primarily financial ratios based on balance sheet and income statement information, and only a few cash flow measures were included.

III. CASH FLOW COMPONENTS

We developed a total cash flow system that had 12 cash flow components (CFC), Gentry, Newbold, and Whitford [1985, 1990].[4] The total cash flow system integrates information from the income statement and changes in balance sheet items between two periods. It provides unique insight concerning management's allocation of resources and the overall performance of the firm. An example of the 12 CFC are presented at the top of Exhibit 1.

Relative cash flow components (CFC*) represent the percentage contribution of each CFC to the total cash flow. The relative cash flow components are determined by dividing each component by the total cash flow. An example of CFC* are presented at the bottom of Exhibit 1. A brief overview of the major components shows the proportion each component contributes to the total cash flow. Exhibit 1 shows that 59.8% of the total inflow came from operations, 16.7% was from net financing, and 9.8% from payables. On the outflow side, which are identified with a minus (-) sign, net investment represented 35.3% of the total outflow, receivables composed 21.6%, inventories 17.6%, and dividends 14.7%.

In Exhibit 2, the CFC* are arranged in a hierarchical order that reflects their economic importance in evaluating the financial health of a firm. Generally, financial and credit analysts use the proposed cash flow hierarchy to evaluate a firm's financial strengths and weaknesses. The hierarchical structure of the CFC* highlights the contribution of each component and the net cash flow available after major inflows and outflows are taken into account. An example of the CFC* hierarchy and the relative net cash flow (NCF*), i.e., the net surplus or deficit cash flow position, is presented in Exhibit 2. This example is based on research findings of Gentry, Newbold, and Whitford [1990].

Exhibit 2 shows 92% of Company A's cash inflows originate from operations (NOF*). After deducting from NOF* the major outflows for investment--NIF* (-45%), and changes in net working capital (-13%), the remaining cash flow surplus represents 34% of the total. The two major outflows associated with the costs of external financial capital are interest expense, [fixed coverage expenditures (FCE*)] and dividends (DIV*). After deducting the FCE*, the surplus cash flow available for dividends (DIV*) is 32%. DIV* consume 12% of total outflows, which leaves a net cash flow surplus of 20%. The surplus cash is used to retire debt (-10%) and invest in marketable securities (-10%). In contrast Company D, an example of a distressed company, has 15% of its cash inflow coming from operations. After deducting cash outflows of 18% for total investment, NIF* being 15% and a net reduction in working capital is 3%, Company D has a deficit cash flow equal to -3% of the total cash flow. The FCE* represents 20% of the total outflow, which leaves a -23% to pay DIV*. DIV* adds an additional 2% to total outflow. The -25% represents a net cash flow deficit and shows that Company D has used all of its operating and working capital cash inflows plus an additional 25% to cover the outflows for investment, dividends and fixed coverage expenditures. Exhibit 2 also shows the deficit was offset by an increase in financing and a decrease in net other assets and liabilities.

Exhibit 2 illustrates several basic concepts that exist between the net cash flow surplus or deficit and levels of risk. First, as the percentage of cash inflows from net operations declines, the net cash flow surplus becomes smaller or the deficit becomes larger. Second, as the net cash flow surplus declines or the net cash flow deficit increases, a firm's financial risk increases. For example, Firm A has the highest net cash flow surplus and it has the lowest financial risk. In contrast, Firm D had the largest net cash flow deficit and it has the highest financial risk. Third, as the relative cash inflow from operations (NOF*) decreases, the relative cash outflow to capital investment decreases. In turn, as the relative cash outflow for interest expense (FCE*) increases, the outflow for DIV* decreases. The pattern of the interrelationships among the key cash flow components is closely associated with the financial health of a firm.

IV. THE ID3 METHOD: INDUCTION OF DECISION TREES[5]

ID3 is an inductive learning program designed by Quinlan [1986], that is based on the original work of Hunt [1966]. ID3 uses data cases of a known class described in terms of a fixed set of attributes, and produces a decision tree of these attributes that correctly classifies the given cases.

The induction of decision trees is based on the process of dividing a group of training examples by the value of a selected attribute where the examples in a group belong to the same class. Thus, an important step in building the inductive tree is selecting the best attribute to branch. ID3 employs entropy as a yardstick for this selection, Shannon [1948, 1951].

The concept of entropy is used in information theory to measure the amount of information transmitted by an information source, based on the number of bits needed to encode all possible messages in an optimal coding. Let $x_1,\ldots,x_n$ be n possible messages occurring with probability $q(x_1),\ldots,q(x_n)$, where $\sum_{i=1}^{n} q(x_i) = 1$. The expected information content, i.e., the entropy, conveyed by the messages is

$$H(x) = -\sum_{i=1}^{n} q(x_i)\log_2 q(x_i)$$

H(x) can be interpreted as the amount of information needed to decode the messages. The higher H(x), the more uncertainty about the content of the message, Ash [1965]. Suppose an attribute may take on two possible values, x and x′, with p(x) = p and p(x′) = 1-p. The entropy in the case of two possible types of messages with probabilities p and 1-p is

$$H = -p \log_2 p - (1-p) \log_2 (1-p) .$$

The function H is symmetric at p and 1-p and maximized at p = 0.5. When p = 1 or 0, there is no uncertainty and hence the entropy H = 0. When p = 1/2, there exists maximum uncertainty as to whether x or x′ will occur, and hence H has the maximum value.

A decision tree for classifying data cases can be regarded as an information source, or a decoder, that generates a message indicating the classification for a given data case. When a node of the tree contains only data cases of the same class, the entropy is equal to zero, which means that the classification decision is defined for the data cases belonging to that node. The induction of the decision tree is thus the process of selecting an attribute to branch that results in the maximal reduction of entropy--which can also be viewed as a process of minimizing uncertainties or maximizing information gains. An example of measuring entropy is presented in Appendix A.

The decision tree is generated by, starting with a root node, progressively selecting attributes to branch the tree. At each iteration of generating the decision tree, ID3 examines all candidate attributes and chooses the attribute that can maximize the amount of information gained. A top-down divide-and-conquer approach is used for *specializing* during the process of induction, i.e., the process subdivides and assigns the cases of the training set at a node into two or more smaller subsets. Therefore, the larger the tree, the more it is specialized to specific case subsets. Consequently, *generalization* of a decision tree, which is the inverse of specialization, can be achieved by *pruning* the tree from the bottom-up based on some evaluating criterion. This is the case for the C4.5 version of ID3 program used in this study.

Examples of the criteria that are used are: (1) the complexity of the resulting tree, (2) the number of terminal nodes in the tree, Breiman, et al., [1984], and (3) the number of cases present at a node that represent each of the classes. The last case occurs because the number of cases decreases as we traverse along a branch of a decision tree from top to bottom, which leads to insignificant splitting due to inadequate sample sizes. In reducing the complexity of decision trees by pruning, Breiman, et al. [1984] used the number of terminal nodes and the misclassification cost of the generated tree as a measure of computational complexity.[6]

Pruning not only reduces the size of a decision tree, it decreases the effect of noise in the data. Real-world data used in a training sample contain a reasonable amount of noise. The negative effect of noise increases from the root of the tree downward because the leaf nodes contain fewer number of cases per represented class. Pruning helps to reduce the propagation of the error by maintaining the number of instances per class at any given node at a desired level, which reduces the effect of noise. Pruning a tree may increase the number of classification errors made on the training (model development) data, but should decrease the error rate on the independent test (holdout) data, Mingers [1989, p. 228].

V. DATA

The acquisition of the data started when a large regional bank agreed to share balance sheet and income statement data for a sample of industrial companies with whom they had an ongoing lending relationship. The bank provided annual data for 44 companies for the period 1985-1986 and for a separate set of 103 companies in 1986-1987. The creation of the high yield (junk) bond market in the mid-1980s, plus the internal expansion in the use of commercial paper to finance short-term corporate needs caused a significant change in the types of firms seeking bank credit. The traditional lending relationships with large industrial companies no longer existed. Regional banks sought new customers in the so-called middle market. The 147 sample companies included in this study were industrials, and they fell in this middle market category. The sales of the sample firms were generally between $50 and $100 million. A few had sales between $100 and $200 million, but the mean was approximately $72 million. None of these companies were listed on a stock exchange, and none were included in the COMPUSTAT data files.

In addition to the accounting information, the bank provided a loan risk ranking for each firm. Also the bank provided three pieces of qualitative information regarding the liquidity status of the loan's collateral and indicated if the loan was secured or unsecured, as well as guaranteed or not guaranteed. A fundamental difference between the loan risk studies of DK and MPW and our study is in the risk class distribution of the sample companies. As indicated earlier 78 percent of DK's sample companies and 90 percent of MPW's sample companies were classified as Category I loans, which is an Office of the Comptroller of Currency (OCC) classification code that represents a normally acceptable banking risk. In the DK and MPW studies these loans were classified as having the single rating of Category I. For these two studies the respective models correctly classified 85 percent and 90 percent of these Category I companies. In our study the bank indicated all of the 147 sample companies were classified as being in Category I. Additionally, the bank assigned these sample companies into five separate risk classes, where a ranking of one was the lowest risk level and a ranking of five was the highest. A contribution of this study is placing a company in one of five classes within an aggregate "current" risk class, which is significantly more difficult than predicting if it should be included in the one large Category I classification.

VI. INDUCTIVE LEARNING ANALYSIS

The inductive learning system is based on the structure of the variables that existed among the 72 companies in the training sample. This structure is used to test a holdout sample of 75 companies. The training sample uses the 12 relative cash flow components, TCF/TA and the three qualitative measures. The means and standard deviations for each of the 12 cash flow components and TCF/TA are presented in Exhibit 3. The objective is to use these 16 variables in an inductive learning system to classify and predict the bank's five loan rating classes. The entropy method selects the variables according to the amount of information added at each level of the decision tree.

The hierarchy of the relative cash flow components (CFC^*) provides a theoretical foundation for hypothesizing the structure of these components in a decision tree. That is, the net operating cash flows (NOF^*) would be the root node followed closely, but not in any specific order, by the most important working capital variables (ΔARF^*, $\Delta INVF^*$, $\Delta OCAF^*$, ΔAPF^*, $\Delta OCLF^*$), fixed coverage expenditures (FCE^*), net investment (NIF^*), and dividends (DIV^*). We do not have a theory to hypothesize where the qualitative variables will appear in the structure.

In testing the accuracy and stability of the C4.5 inductive learning system, multiple trees were generated. Each tree had a unique structure and used a different combination of attributes. The decision tree in Figure 1 is presented as a reasonable proxy of the various trees generated by the inductive learning system. From the 16 variables, the entropy technique determined dividends (DIV^*) was the most discriminating variable. The left branch coming from DIV* shows that 16 of the companies paid out 12.3% or more of total cash outflow to dividends, while the right branch indicates 56 of the companies paid out less than 12.3% of the total outflow in dividends.

In summary, six of the 16 companies on the left branch were correctly classified as 3s, seven were correctly classified as 2s and two companies were correctly placed in the 1 class. One company originally in class 1 was misclassified as a 2. Thus, after four levels of linear and sequentially related information, the inductive tree system classified correctly 15 companies and misclassified one company.

The right-hand branch coming from the root node in Figure 1 indicates that these 56 companies distributed a smaller proportion of their cash outflow to dividends than did the 16 companies on the left-hand branch. Further, the right-hand side of the tree shows there are a different set of attributes used in classifying the ratings of the bank's loans vis-a-vis the left-hand side of the tree. The right side of the tree is longer and the branches are more complex. That is, the tree splits into subbranches at two separate nodes, ΔAPF* and ΔINVF*. By the fourth level only eight of the 56 companies were correctly classified and two companies originally in class 3 were misclassified as 2s. At the fourth level for both branches of the tree, approximately one-third (23/72) of the companies were correctly classified and three were misclassified.

The analysis reveals that 95% (68/72) of the loan risk ratings in the training example were classified correctly by the inductive tree system.[7] A holdout or test sample of 75 companies was used to test the predictive accuracy of the inductive learning system. The test results showed that 56% (42/75) of the loan risk ratings were predicted correctly by the inductive learning system.[8] Furthermore, to test the classification stability of the inductive learning system several pruning confidence levels were used. It was found that the classification accuracy remained constant at 56.9% for pruning confidence levels between .01 and .10.

The most insightful output generated by the C4.5 model was that each of the 75 testing companies received a 3 rating. That is, for the 75 holdout companies the inductive learning system could not detect any difference in their loan risk ratings. This surprising finding highlights the challenge involved in determining loan ratings of high risk companies.

We reported this finding to the bank officers, who, in turn, indicated that in their judgment, there was a distinct difference in the risk ratings between the low risk loans, the 1s and 2s, and the high risk loans, the 4s and 5s. Thus, the remaining challenge was to determine if the loans rated 3 by the bank could be separated into two separate risk groups, one composed of low risk firms and the other representing a group of high risk firms. The bank officers indicated they would find it extremely valuable to know which of the loans they rated as 3s more resembled the low risk rating class, the 1s and the 2s, or, alternatively which of the 3s are more closely aligned with the high risk group, the 4s and 5s. A dynamic updating process developed by Tessmer [1992] makes it possible to classify the loans rated as 3s into either a low or a high risk class.

The first step is to use the inductive learning system to determine the differences between the low and the high risk class loans. There are 36 low risk loans rated 1 or 2 and 28 high risk loans rated 4 or 5. The cash flow and qualitative information for these 64 companies are utilized in the C4.5 system. The result is an induced decision tree composed of low and high risk companies. Because each induced tree can have a unique structure, Tessmer [1992, pp. 12-15], a jackknife procedure was used to repeat the experiment 64 times.

The result is a final global tree shown in Figure 2, Tessmer [1992, pp. 12-15]. The final global tree is a composite of the 64 original trees. The global tree reduces noise and overfitting effects that are present in the original trees. Figure 2 retains the most frequently appearing attributes in their most likely position in the original trees.

The most important attribute in Figure 2 is the dividend (DIV) component. The global tree selects only four attributes--dividends, net financing flows, secured/unsecured classification and net operating cash flows--to classify the loan ratings of the 64 companies into a high or a low risk rating. The average prediction accuracy among the 64 original trees was 89 percent. The inductive learning system correctly predicted 35 of 36 low risk loans and 22 of 28 high risk loans. That is 31 of the low risk loans were correctly classified because they paid a dividend. In addition the remaining four low risk loans had the following characteristics: they did not pay a dividend, the financing flows were less than 50 percent of the total inflows, the loans were unsecured and, finally, 50 percent or more of their total cash inflows were generated from operations.

The structure of the decision tree for the high rated loans shows that five did not pay a dividend and external financing sources composed 50 percent or more of their total cash inflows. In addition to these two attributes, 16 of the high risk companies had a secured contractual arrangement with the bank. The remaining high risk company did not pay a dividend, received less than 50 percent of total cash inflow from financing, was an unsecured credit and had net operating cash inflows of 50 percent or less of the total cash inflows.

The next step of the dynamic updating process is to use the induced global tree in Figure 2 to classify the 84 loans that are rated 3s by the bank staff into either a low or a high risk rating class. Using the Figure 2 decision tree, 47 of the 84 companies were reclassified as having attributes that more closely resembled the low risk class. The remaining 37 companies were found to have attributes that resembled the high risk class. The reclassification of the 3s is illustrated in Figure 3.

The 147 original trees have been reclassified into 83 low risk examples [11 (class 1) + 25 (class 2) + 47 (class 3)] and 65 high risk examples [24 (class 4) + 4 (class 5) + 37 (class 3)]. The jackknife method was used to induce 147 original trees based on the new reclassification. A final global tree is a composite of these 147 original trees, Tessmer [1992], and it is presented in Figure 4. The global tree is composed of seven cash flow components and one qualitative measure, the secured/unsecured status of each loan. The jackknife results that underlie the global tree resulted in a mean prediction accuracy rate of 92 percent across the 147 decision trees. The global tree in Figure 4 generated a prediction accuracy rate of 88.4 percent. These findings indicate a high level of predictability associated with the global tree approach.

The root node was the dividend component (DIV^*). By discovering that a company paid a dividend, the C4.5 system correctly predicted a low risk rating for 68 companies and misclassified eight high risk companies as being low risk. The second most important attribute was the secured/unsecured status of the loan. It split the tree into two major subbranches. The branch on the left used three additional cash flow components to classify correctly the remaining low risk loans and approximately one-fourth of the high risk loans. The branch on the right correctly classified all of the remaining high risk loans.

The structure of the low risk loans in Figure 4 was that (1) they paid a dividend (DIV^*); (2) the loan was unsecured and the net operating cash flow (NOF^*) was greater than 60 percent of the total inflow; (3) the change in other assets and liabilities flows ($\Delta OA\&LF^*$) was greater than 10 percent of the total outflow; and finally capital expenditures (NIF^*) were greater than 15 percent of total outflow. Approximately one-fourth of the high risk loans had the same structure except that capital expenditures (NIF^*) were less than 15 percent of total outflow.

In contrast, the structure of approximately three-fourths of the high risk loans was that (1) they did not pay a dividend; (2) the loans were secured; (3) net financing flows composed more than 45 percent of the total cash inflow; (4) the change in other current liabilities (OCL^*) was in a range between 20 percent of total cash outflow and 25 percent of total cash inflow; and (5) the change in inventory was less than 20 percent of total outflow.

Observations Concerning Inductive Learning

Several significant observations evolve from the inductive learning analysis. Earlier, a hierarchy of cash flow components was developed, and it was hypothesized that the net operating cash flow component (NOF^*) would be the root node in the induced decision trees. However, the results show that DIV^* was the root node, which makes it the most discriminating cash flow component in classifying loan risk. This finding supports previous empirical test results that predicted bond ratings and bankruptcy, Gentry, Newbold and Whitford [1985a, 1985b, 1988]. Why isn't NOF^* the root node as hypothesized? Although it is conjecture, DIV^* serves as a proxy for NOF^*. The surplus cash flow available for paying dividends is dependent on a firm's operating performance in executing its strategic plans. Although there are several operating and strategic decisions and actions responsible for generating a surplus net cash flow, the NOF^* is the theoretical foundation for making a surplus cash flow available for paying dividends. Thus, without the availability of relatively large cash flows from operations, a cash payment to DIV^* is difficult to accomplish. In essence, DIV^* reflects a firm's dividend policy, but more importantly it provides a signal to the financial markets that the firm has the cash available to pay dividends to its shareholders.[9]

Tree induction reveals several characteristics of the cash flow data being analyzed. The presence of only a few nodes on the tree signals that distinct information patterns exist which make it possible to discriminate among the risk classes. Furthermore, a small linear tree indicates that a straight sequence of a few variables can easily determine a firm's credit risk rating. Another dimension of the tree induction process is that the most discriminating and important variables are close to the root node. Likewise the value added by the components in the lower levels of the tree is less than the value contributed by the components closer to the root of the tree. When several components are used to determine a risk rating, it indicates the complexity of the information system needed to differentiate the subtle risks that exist in the data.

The inductive learning system assigned all 75 of the test (holdout) companies a loan risk classification rating of 3. The result was a predictive accuracy of 56 percent. Additionally, we observed that the induction of multiple trees resulted in a modest instability in the structure of the induced trees and also in the position of the attributes in the trees. We concluded there was a need to improve the stability of the induced decision trees and the predictive accuracy of the system. To achieve these two critical objectives, we set out to determine if we could use a dynamic updating process developed by Tessmer [1992]. First, we learned that the lending officers were quite interested in being able to differentiate between the low and the high risk loans. Therefore, we were able to use the dynamic updating process in the inductive learning system to classify the loans into either a low or a high risk category. The process used a jackknife procedure to generate a global decision tree that was a composite of the decision tree structure and the position of the attributes in the tree. The global tree reduces the noise and complexity that exists in the original trees. Thus, it improves the stability of the induced tree and substantially improves its predictive accuracy. Finally, the simplified structure of the global tree greatly enhances the insights of credit analysts in interpreting the components that underlie loan risk ratings.

VII. PROBIT ANALYSIS

As a final test, the same data were used in a polytomous probit model, except the change in cash variable was omitted. The results in Exhibit 4 show five of the probit coefficients are statistically significant at the 5 percent level or higher; these include NOF^*, FCE^*, DIV^* and liquidity of collateral as well as the intercept term. The training set data had a classification accuracy of 65.3 percent (47/72). The probit coefficients in Exhibit 4 were used to predict the loan risk ratings of the 75 companies in the holdout sample. The probit model correctly predicted 56 percent (42/75) of the bank loan risk ratings for the holdout sample, as shown in Exhibit 5. The probit predictive results are almost identical to the inductive learning results.[10]

VIII. CONCLUSIONS

The primary contribution of this paper is the development of an inductive learning system for classifying loan risk. In classifying loan risk the system generates a decision tree structure of the credit rating information. The hierarchical structure of the inductive decision tree adds valuable insight concerning the subtle differences and nuances that exist among risky firms. The decision tree structure provides a tool for a rich and deep interpretation of the components that produce the loan risk ratings.

The inductive trees generated by the system used four to eight components to determine the risk rating. The cash flow components selected by the induction trees were quite similar to the hypothesized cash flow component hierarchy. When multiple trees were developed, it was apparent that the structure of the original trees were modestly unstable as were the positions of the attributes in the trees. Thus, a dynamic updating procedure was used to stabilize the structure of the tree as well as to improve its predictive accuracy. Additionally, the simplified structure of the resulting global decision tree greatly enhances the insights of the credit analysts in interpreting the components that underlie the loan risk ratings. Other significant observations concerning the global tree were that dividends were the most discriminating relative cash flow component and secured/unsecured status of the loan was the second most important attribute. The inductive learning system associated with the global tree was able to classify the loans into low and high risk with a very high degree of accuracy. In essence, the global tree results indicate that loan risk classification is highly predictable and provides a valuable tool for credit analysts.

FOOTNOTES

[1] The authors are very appreciative of the financial support provided by the Prochnow Educational Foundation and the KPMG Peat Marwick Foundation for sponsoring this research project. Also the authors are grateful for the very capable research assistance of Fuh-Jiun Kuo, Hsing-Yao Chen, Selwyn Piramuthu and Chau Chen Yang in conjunction with this research project.

[2] The dummy variables related to each loan were: (1) unsecured or secured, (2) past due or current in payment, (3) clean audit opinion or not, (4) net loss or net profit and (5) criticized or not criticized by the examiner in the last period.

[3] Haslem and Longbrake [1972] were critical of Orgler's using outside examiner ratings rather than the rating of an insider, such as the lending officer. Also, they objected to the use of past information to explain a current rating, but did not offer an alternative.

[4] The analysis also included a scalar variable, Total Cash Flow/Total Assets (TCF/TA).

[5] The following section is based on a presentation in Shaw, Gentry and Piramuthu [1990].

[6] In creating the original tree, a goodness of split measure determines the attribute at a node. The value of the measure reflects how well the chosen attribute splits the data between classes at that node. A pruning method, called critical value pruning, specifies a critical value, typically between .95 and .9995, and prunes those nodes which do not reach it. However, pruning does not occur if a node further along the branch does reach the critical value. The larger the critical value selected the greater the degree of pruning and the smaller the resulting tree. In practice, a series of pruned trees is generated using increasing critical values, Mingers [1989, pp. 231-232].

[7] Numerous training sets were tested to take into account the degree of uncertainty that exists in the data, Mingers [1989, p. 228]. Mingers indicates this uncertainty may arise from two different sources. The first is mis-measurement which may occur for a variety of reasons, which is referred to as noise. The second source of uncertainty is the occurrence of extraneous factors which are not recorded, which is called residual variation.

[8] Mingers [1989, p. 236] also states there are two important criteria for evaluating a decision tree--size and accuracy. One objective is to minimize the size of the induced decision tree as measured by the number of nodes. In this project the number of nodes was selected because it reflected the number of decision rules contained

in the decision tree. The second objective is accuracy or the predictive ability of a decision tree to classify an independent set of test data. It is measured by the error rate, which is a crude measure because it does not reflect the accuracy of predictions for different classes within the data. In this data set the risk classes were not equally likely and Mingers [1989, p. 237] observes those with few examples are usually predicted badly.

[9] Compared to high risk companies the data show that companies with low risk ratings distribute a higher proportion of their total outflow to dividends. The training model included 17 low risk companies, i.e., companies with a risk rating of either 1 or 2. The mean DIV^* of these 17 companies was 15.09 percent. Only four of these companies did not pay a dividend. In contrast, only two of the 17 high risk companies in the training set paid a dividend, i.e., companies with a risk rating of either 4 or 5. The DIV^* for the two companies was 5.67 percent and 1.33 percent. Additionally, the availability of NOF^* has an impact on DIV^*. In the study 76 percent (13/17) of the low risk companies had a NOF^* greater than 50 percent. However, only 36 percent (5/14) of the high risk companies had a NOF* greater than 50 percent. These findings support the observation made by Miller and Rock [1985] regarding dividend signalling. That is, the best place "to look for signalling may well be among firms falling into adversity, not because they start signalling but because they stop" (p. 1046).

[10] Exhibit 5 indicates that approximately 56 percent (42/75) of the loan ratings in the holdout sample were accurately predicted and an additional 41 percent (31/75) of the predicted ratings were in a cell that was adjacent to the actual rating. Thus 97 percent (73/75) of the predicted ratings are either correct or within one rating class of the actual, where the model's second highest conditional probability classification was the correct rating. To acquire additional insight into the prediction quality Exhibit 5 shows that of the 75 loans in the holdout sample, 55 had ratings that were either a 2 or a 3. Thus, a naive predictor that classified all loans a 3 would be correct or within one rating class of a 2 rating 55 times out of 75. The null hypothesis that our predictor performs at this level can be tested through a chi-square goodness of fit test. Since 73 of the 75 loans were predicted correctly or within one class, the calculated test statistic is $(73-55)^2/55 + (2-18)^2/18 = 20.1$. Compared with tabulated values of the chi-square distribution with one degree of freedom, the null hypothesis that our predictor performs at the level of the naive predictor can be rejected at the 5 percent significance level.

REFERENCES

Altman, E. I., R. B. Avery, R. A. Eisenbeis and J. F. Sinkey, Jr., Application of Classification Techniques in Business, Banking and Finance, Greenwich, CT: JAI Press, Inc., 1981.

Ash, R., Information Theory, New York, NY: John Wiley & Sons, 1965.

Aziz, A. and G. H. Lawson, "Cash Flow Reporting and Financial Distress Models: Testing of Hypotheses," Financial Management, Vol. 19 (Spring 1989), pp. 55-63.

Belkaoui, Ahmed, Industrial Bonds and the Rating Process. Westport, CT: Quorum Books, 1983.

Bernard, V. L. and T. L. Stober, "The Nature and Amount of Information in Cash Flows and Accruals," The Accounting Review, Vol. 65 (October 1989), pp. 624-652.

Bowen, R. M., D. Burgstahler and L. A. Daley, "The Incremental Information Content of Accrual versus Cash Flows," The Accounting Review, Vol. 62 (October 1987), pp. 723-747.

Breiman, L., Freidman, J. H., Olshen, R. A., and Stone, C. J., Classification and Regression Trees, California: Wadsworth Publishing Company, 1984.

Casey, C. J. and N. J. Bartczak, "Cash Flow--It's Not the Bottom Line," Harvard Business Review, Vol. 62 (July-August 1984), pp. 60-66.

____________, "Using Operating Cash Flow Data to Predict Financial Distress: Some Extensions," Journal of Accounting Research, Vol. 23 (Spring 1985), pp. 384-401.

Dambolena, I. G. and J. M. Shulman, "A Primary Rule for Detecting Bankruptcy: Watch the Cash," Financial Analysts Journal, Vol. 44 (September/October 1988), pp. 74-78.

Dietrich, R. and R. Kaplan, "Empirical Analysis of the Commercial Loan Classification Decision," The Accounting Review, Vol. 57 (January 1982), pp. 18-38.

Duchessi, P., H. Shawky and J. P. Seagle, "A Knowledge-Engineered System for Commercial Loan Decisions," Financial Management, Vol. 17 (Autumn 1988), pp. 57-65.

Foster, G., Financial Statement Analysis, Second Edition, Englewood Cliffs, NJ: Prentice-Hall, 1986.

Fridson, M. S. and M. A. Cherry, "Dispersion of Corporate Bond Returns within Quality Classes," Institutional Investor (November 1990), pp. 4-9.

Gentry, J., P. Newbold, and D. T. Whitford, "Classifying Bankrupt Firms with Funds Flow Components," Journal of Accounting Research, Vol. 23 (Spring 1985a), pp. 146-160.

____________, "If Cash Flow's Not the Bottom Line, What Is?" Financial Analysts Journal, Vol. 41 (September/October 1985b), pp. 47-56.

____________, "Predicting Industrial Bond Ratings with a Probit Model and Funds Flow Components," Financial Review, (August 1988), pp. 269-286.

____________, "Profiles of Cash Flow Components," Financial Analysts Journal, Vol. 46 (July/August 1990), pp. 41-48.

Gombola, M. S., M. E. Haskins, J. E. Ketz, and D. D. Williams, "Cash Flow in Bankruptcy Prediction," Financial Management, Vol. 16 (Winter 1987), pp. 55-65.

Haslem, J. and W. Longbrake, "A Credit Scoring Model for Commercial Loans, A Comment," Journal of Money, Credit and Banking, (Spring 1972), pp. 733-734.

Helfert, E., Techniques in Financial Analysis, 5th Edition, Homewood, IL: Richard D. Irwin, 1982.

Horrigan, James D., "The Determination of Long-Term Credit Standing with Financial Ratios," in Empirical Research in Accounting: Selected Studies, Vol. 62 (1966), pp. 44-62.

Hunt, E. B., Concept Learning: An Information Processing Problem, New York, NY: John Wiley & Sons, 1962.

Jensen, M. C., "The Takeover Controversy: Analysis and Evidence," Midland Corporate Finance Journal, Vol. 4 (Summer 1986), pp. 6-32.

Jensen, M. C. and R. S. Ruback, "The Market for Corporate Control: The Scientific Evidence," Journal of Financial Economics, Vol. 11 (April 1983), pp. 5-50.

Kaplan, R. and G. Urwitz, "Statistical Model of Bank Ratings: A Methodological Inquiry," Journal of Business, Vol. 52 (April 1979), pp. 231-261.

Largay, J. A. and C. P. Stickney, "Cash Flows, Ratio Analysis and the W. T. Grant Company Bankruptcy," Financial Analysts Journal, Vol. 36 (July/August 1980), pp. 51-54.

Lehn, K. and A. Poulsen, "Free Cash Flow and Stockholder Gains in Going Private Transactions," Journal of Finance, Vol. 44 (July 1989), pp. 771-787.

Livnat, J. and P. Zarowin, "The Incremental Information Content of Cash-Flow Components," Journal of Accounting and Economics, Vol. 3 (1990), pp. 25-46.

Marais, J. L., J. Patell and M. Wolfson, "The Experimental Design of Classification Models: An Application of Recursive Partitioning to Commercial Loan Classifications," Journal of Accounting Research, Vol. 22 (Supplement 1984), pp. 87-114.

Miller, M. and K. Rock, "Dividend Policy under Asymmetric Information," Journal of Finance, Vol. 40 (September 1985), pp. 1031-1051.

Mingers, J., "An Empirical Comparison of Pruning Methods for Decision Tree Induction," Machine Learning, Vol. 4 (1989), pp. 227-243.

Neill, J. D., T. F. Schaefer, P. R. Bahnson and M. E. Bradbury, "The Usefulness of Cash Flow Data: A Review and Synthesis," Journal of Accounting Literature, Vol. 10 (!991), pp. 117-150.

Orgler, Y. E., "A Credit Scoring Model for Commercial Loans," Journal of Money, Credit and Banking, (November 1970), pp. 435-445.

Pinches, George E. and Kent A. Mingo, "A Multivariate Analysis of Industrial Bond Ratings," Journal of Finance, Vol. 23 (March 1973), pp. 1-18.

Porter, M. E., Competitive Strategy, New York: Free Press, 1980.

Quinlan, J. R., "Induction of Decision Trees," Machine Learning, Vol. 1 (1986), pp. 81-106.

Rayburn, J., "The Association of Operating Cash Flow and Accruals with Security Returns," Journal of Accounting Research, Vol. 24 (Supplement 1986), pp. 112-137.

Shannon, C. E., "A Mathematical Theory of Communication," Bell System Technology Journal, Vol. 27 (July 1948), pp. 379-423, (October 1948), pp. 623-656.

____________, "Prediction and Entropy of Printed English," Bell System Technology Journal, Vol. 30 (January 1951), pp. 50-65.

Shaw, M. J. and J. A. Gentry, "Using An Expert System with Inductive Learning to Evaluate Business Loans," Financial Management, Vol. 17 (Autumn 1988), pp. 45-56.

Shaw, M. J., J. A. Gentry and S. Piramuthu, "Inductive Learning Methods for Knowledge-Based Decision Support: A Comparative Analysis," Computer Science in Economics and Management, Vol. 3 (1990), pp. 147-165.

Snyder, C. L., Jr., "Credit Score Theory," The Loan Pricing Report, Vol. 5, No. 9 (October 1990), pp. 4-10.

Srinivasan, V. and Y. H. Kim, "Designing Expert Financial Systems: A Case of Corporate Credit Management," Financial Management, Vol. 17 (Autumn 1988), pp. 32-44.

Tessmer, A. C., "New Dimensions of Inductive Learning for Credit Risk Analysis," Working Paper, College of Commerce and Business Administration, University of Illinois, 1992, 27 pages.

von Stein, J. H. and W. Ziegler, "The Prognosis and Surveillance of Risks from Commercial Credit Borrowers," Journal of Banking and Finance, Vol. 8 (1984), pp. 249-268.

Wilson, G. P., "The Incremental Information Content of the Accrual and Funds Components of Earnings After Controlling for Earnings," The Accounting Review, Vol. 62 (April 1987), pp. 293-322.

____________, "The Relative Information Content of Accruals and Cash Flows: Combined Evidence at the Earnings Announcement and Annual Report Release Date," Journal of Accounting Research, Vol. 24 (Supplement 1986), pp. 165-200.

EXHIBIT 1

AN EXAMPLE OF CASH FLOW COMPONENTS (CFC)

CASH INFLOWS (+)		CASH OUTFLOWS (-)	
NET OPERATING	$1220	Δ RECEIVABLES	$440
Δ OTHER C.A.	40	Δ INVENTORY	360
Δ PAYABLES	200	FIXED COVERAGE EXP.	180
Δ OTHER C.L.	100	NET INVESTMENT	720
Δ NET FINANCIAL	340	DIVIDENDS	300
Δ CASH M.S.	140	Δ NET OTHER A & L	40
TOTAL CASH FLOW (+)	$2040	TOTAL CASH FLOW (-)	$2040

AN EXAMPLE OF RELATIVE CASH FLOW COMPONENTS (CFC*)[1]

CASH INFLOWS (+)	% OF TOTAL CASH FLOW (+)	CASH OUTFLOWS (-)	% OF TOTAL CASH FLOW (-)
NET OPERATING*	59.8	Δ RECEIVABLES*	21.6
Δ OTHER C.A.*	2.0	Δ INVENTORY*	17.6
Δ PAYABLES*	9.8	FIXED COVERAGE EXP.*	8.8
Δ OTHER C.L.*	4.9	NET INVESTMENT*	35.3
Δ NET FINANCING*	16.7	DIVIDENDS*	14.7
Δ CASH M.S.*	6.8	Δ NET OTHER A & L*	2.0
	100%		100%

[1] $\frac{\text{CASH FLOW COMPONENT}}{\text{TOTAL CASH FLOW}} = \text{RELATIVE CASH FLOW COMPONENT}$

*Indicates relative cash flow as opposed to actual cash flow.

EXHIBIT 2

AN EXAMPLE OF THE HIERARCHY OF RELATIVE CASH FLOW COMPONENTS UNDER VARIOUS RISK CONDITIONS

	Company			
	Lowest Credit Risk			Highest Credit Risk
Relative Cash Flow Components (CFC*)	A	B	C	D
Net Operating (NOF*)	92%	70%	57%	15%
ΔAR*	-9	-15	-22	30
ΔINV*	-11	-17	-18	25
ΔOCA*	-1	-3	2	10
ΔAP*	7	15	17	-43
ΔOCL*	1	8	9	-25
Net Investment (NIF*)	-45	-38	-30	-15
Surplus or Deficit after Investment Expenditures	34	20	15	-3
Fixed Coverage Exp. (FCE*)	-2	-6	-9	-20
Surplus or Deficit available for dividends	32	14	6	-23
Dividends (DIV*)	-12	-14	-15	-2
Net Cash Flow Surplus or Deficit (NCF*)	20%	0%	-9%	-25%
ΔNet Financing (ΔNF*)	-10	7	10	22
ΔNet Other A & L (ΔNOA&L*)	0	0	-6	3
ΔCash & M.S. (ΔCash*)	-10	-7	5	0
CFC* After All Cash Flows	0	0	0	0

EXHIBIT 3

MEANS AND STANDARD DEVIATIONS OF THE RELATIVE CASH FLOW COMPONENTS (CFC*), 1986-1987

Cash Flow Component	Mean	S.D.
Operating (NOF*)	.4770	.2538
ΔReceivables (ΔARF*)	-.0861	.2010
ΔInventories (ΔINVF*)	-.0425	.1997
ΔOther CA (ΔOCAF*)	.0050	.0990
ΔPayables (ΔAPF*)	.0588	.1606
ΔOther CL (ΔOCL*)	.0258	.1249
ΔOther A & L (ΔNOA&L*)	-.0501	.2161
ΔFinancing (ΔNFF*)	-.0479	.3183
Fixed Coverage (FCE*)	-.1014	.0927
Investment (NIF*)	-.2459	.1990
Dividend (DIV*)	-.0731	.1175
TCF/TA	.2904	.1425
N	147	

EXHIBIT 4

PROBIT COEFFICIENTS FOR RELATIVE CASH FLOW COMPONENTS AND QUALITATIVE FACTORS FOR THE PREDICTION OF LOAN RISK RATINGS, 1986 AND 1987 COMBINED

Relative Cash Flow Component	Probit Coefficients
Constant	3.156***
Operating (NOF*)	-3.348***
ΔReceivables (ΔARF*)	-0.861
ΔInventories (ΔINVF*)	-1.375
ΔOther CA (ΔOCAF*)	0.513
ΔPayables (ΔAPF*)	-0.361
ΔOther CL (ΔOCLF*)	-0.434
ΔOther A&L (ΔOA&L*)	-1.450
ΔFinancing (ΔNFF*)	-1.233
Fixed Coverage Expenditures (FCE*)	-4.726**
Investments (NIF*)	-1.274
Dividends (DIV*)	3.689**
TCF/TA	-2.171
Dummy Variables	
Secured/Unsecured	0.639
Guarantee/No Guarantee	-0.179
Liquidity of Collateral	2.475**
n	72

**Significant at .05 level of confidence.
***Significant at .01 level of confidence.

EXHIBIT 5

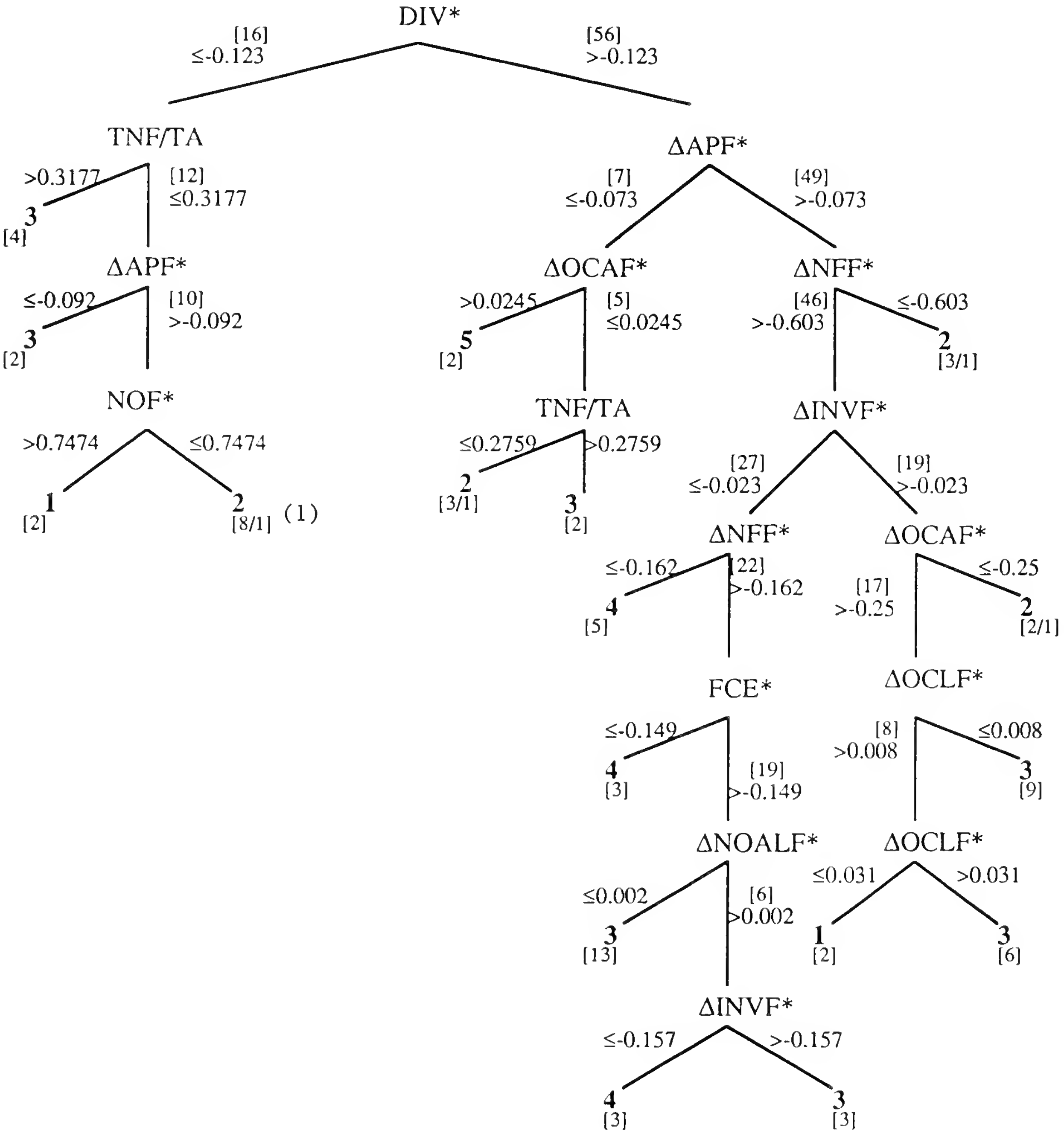
PREDICTION OF LOAN RISK RATINGS IN THE HOLDOUT SAMPLE WITH CASH FLOW COMPONENTS AND QUALITATIVE FACTORS, 1986-1987

Bank	Predicted Ratings					
Ratings	1	2	3	4	5	TOTAL
1	2	3	1			6
2	3	4	6			13
3	1	2	31	8		42
4			7	5		12
5			1	1		2
TOTAL	6	9	46	14	0	75

56 percent of bank loan risk ratings are predicted correctly.

FIGURE 1

Inductive Decision Tree with Five Rating Classes

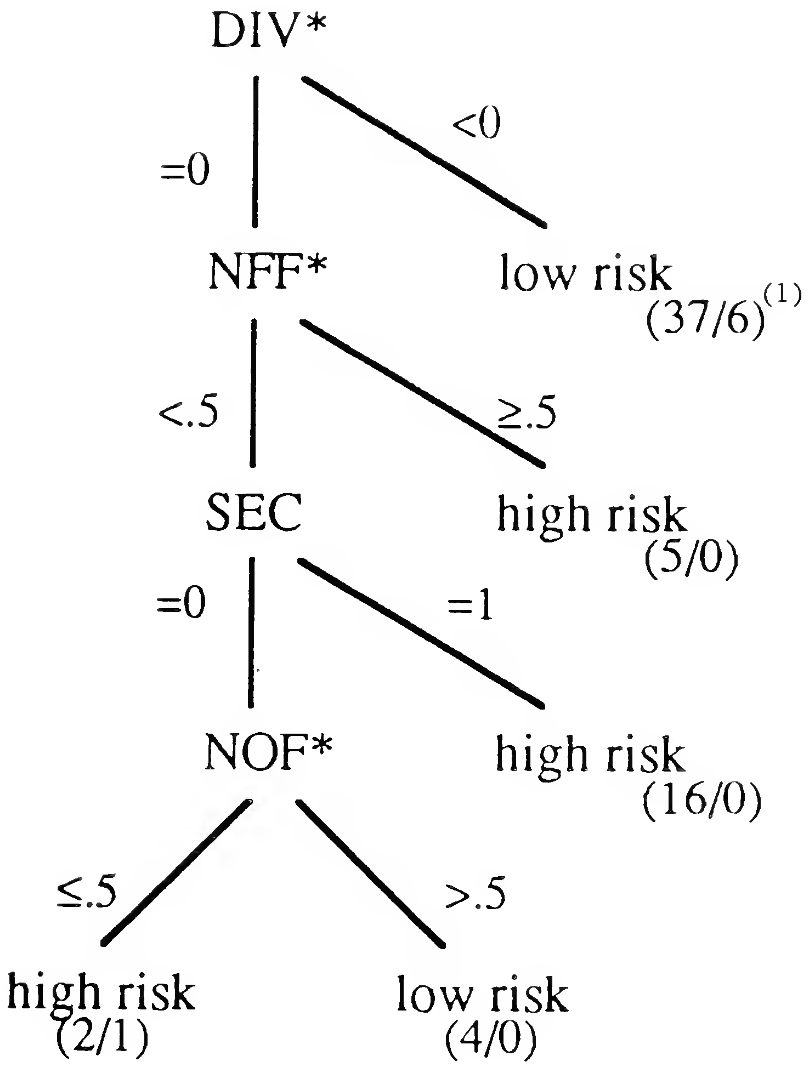


72 training examples - all 15 variables - 5 classes - (PCL = 0.5)

(1) 8 companies are classified by the branch, 1 of them is misclassified, i.e., did not belong to class 2.

FIGURE 2

Global Tree of the 36 Low Risk Loans and the 28 High Risk Loans

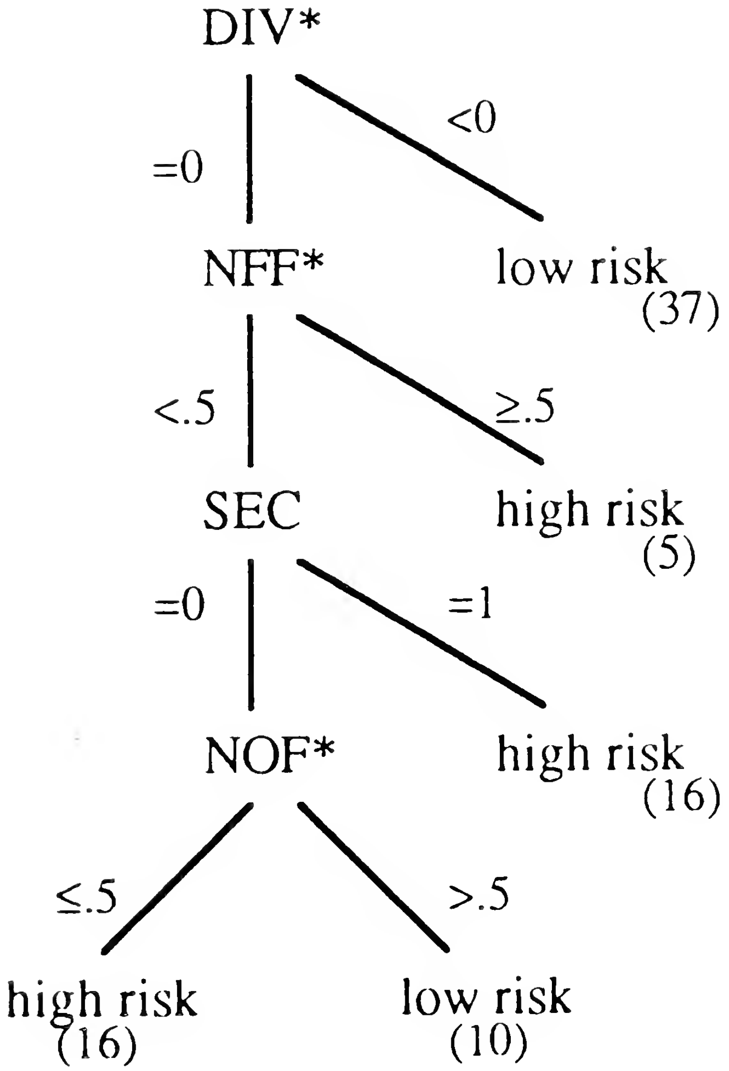


An 87.5 percent Mean Prediction Accuracy (56/64).

[1] (total number of companies classified for the branch/number of companies incorrectly classified for the branch)

FIGURE 3

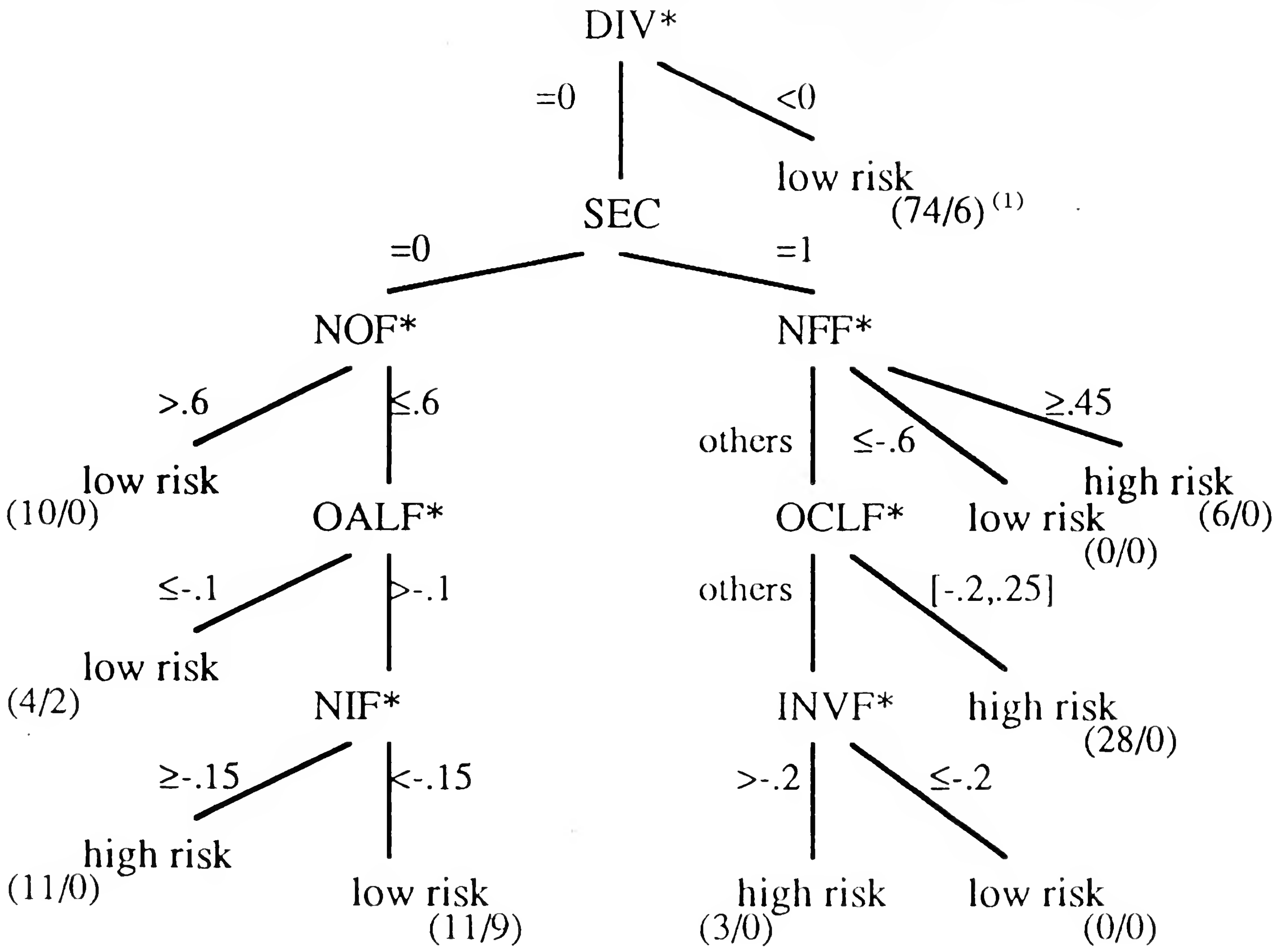
Global Tree of the 84 Loans that were Rated 3s by the Bank Staff



47 reclassified to the low risk class
37 reclassified to the high risk class

Figure 4

Global Tree of 147 Original Trees



An 88.4 percent prediction accuracy (130/147)

[1] (total number of companies classified for the branch/number of companies incorrectly classified for the branch)

APPENDIX A

MEASURING ENTROPY

A simplified loan risk rating training sample is presented in Exhibit 6. It is used to illustrate the operation of the ID3 algorithms. Exhibit 6 contains a sample of 10 firms which are rated as being either a good credit risk or bad credit risk. Only two ratings are used in order to simplify the example. The data for three of the most important relative cash flow components are selected--net operating (NOF*), net investment (NIF*) and dividends (DIV*). The values for these attributes are found in Exhibit 6.

In this example, there are six good credit risk firms and four with a bad credit risk. The probabilities of these events can be estimated by using the relative frequencies. If p is the probability of occurrence of a good credit rating, then p = 0.6 and the probability of a bad credit risk rating is 1 - p = 0.4, as shown in Figure 5.

The total expected information content of this decision is equal to the amount of entropy. It is to be denoted by the entropy (H), or the information content of the tree. Then,

$$H = -0.6 \log_2 0.6 - 0.4 \log_2 0.4 = 0.97. \tag{1}$$

To determine the decision tree, each attribute (variable) must be evaluated as to its appropriateness as a discriminating variable. First, the relative cash outflow going to dividends (DIV*) is tested. Figure 6 provides the information used in ID3. The data are based on the training example in Exhibit 6. When DIV* is low, the amount of entropy associated with the subtree is

$$H = -0.6 \log_2 0.6 - 0.4 \log_2 0.4 = 0.97. \tag{2}$$

When DIV* is high, the entropy of the subtree is also 0.97. Therefore, if the tree is split on DIV*, the expected entropy ($\overline{H}$) after the split is

$$\overline{H} = 0.5 * 0.97 + 0.5 * 0.97 = 0.97. \tag{3}$$

Hence, the amount of information gained by splitting on DIV*, which is the reduction in entropy by the split, is

$$0.97 - 0.97 = 0. \tag{4}$$

The second variable to be tested is the relative net operating cash flow (NOF*), which is shown in Figure 1. When NOF* is small, the entropy of the subtree is

$$-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1.00. \tag{5}$$

When NOF* is medium or large, the entropy of subtree is zero, which implies that there is no uncertainty. Thus, the expected entropy after splitting on NOF* is

$$\bar{H} = 0.4 * 1.00 + 0.2 * 0 + 0.4 * 0 = 0.40. \tag{6}$$

Therefore, the amount of information gained by using NOF* as a node is

$$0.97 - 0.40 = 0.57. \tag{7}$$

The third variable to be tested is relative net investment (NIF*) which is shown in Figure 8. When NIF* is low, the entropy of the subtree is

$$H = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1.00. \tag{8}$$

When NIF* is high, the entropy of the subtree is

$$H = -0.75 \log_2 0.75 - 0.25 \log_2 0.25 = 0.81. \tag{9}$$

Thus, the expected entropy after splitting on NIF* is

$$\bar{H} = 0.6 * 1.00 + 0.4 * 0.81 = 0.92. \tag{10}$$

Hence, the amount of information gained by using NIF* as a node is

$$0.97 - 0.92 = 0.05. \tag{11}$$

The largest amount of information gain is obtained by using NOF*. In other words, NOF* provides the largest reduction of uncertainties with respect to analyzing financial failure. Hence, NOF* is chosen as the root node of the tree. If NOF* is used as the root node, there exists uncertainty only when NOF* is small. Again, DIV* is tested by the same procedure, as shown in Figure 9. When DIV* is low, the amount of entropy associated with the subtree is

$$H = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1.00. \tag{12}$$

When DIV* is high, the same amount of entropy is obtained. Therefore, if the tree is split on DIV*, the expected information content after the split is

$$\overline{H} = 0.4 * [0.5 * 1.00 + 0.5 * 1.00] = 0.4. \tag{13}$$

Hence, the amount of information obtained by splitting on DIV* is

$$0.4 - 0.4 = 0.0 \tag{14}$$

which means that DIV* does not help to gain information. NIF* is then tested as shown in Figure 10. When NIF* is low or high, the amount of entropy associated with subtree is

$$H = -0.0 \log_2 0.0 - 1 \log_2 1.0 = 0.0. \tag{15}$$

Thus, the expected information content after the split is

$$\overline{H} = 0.4 * [0.5 * 0.0 + 0.5 * 0.0] = 0.0. \tag{16}$$

Hence the amount of information obtained by splitting on NIF* is

$$0.4 - 0.0 = 0.4.$$

Therefore, NIF* is selected as second node and the learning process is completed as each leaf contains companies belonging to a single class, as shown in Figure 11.

EXHIBIT 6

LOAN RISK RATING TRAINING EXAMPLE

Firm	Dividend (DIV*)	Operating (NOF*)	Investment (NIF*)	Credit Risk Rating
A	low	small	low	Bad
B	low	small	high	Good
C	low	medium	high	Bad
D	low	large	low	Good
E	low	large	low	Good
F	high	small	low	Bad
G	high	large	high	Good
H	high	small	high	Good
I	high	medium	low	Bad
J	high	large	low	Good

FIGURE 5

Initial Decision Tree

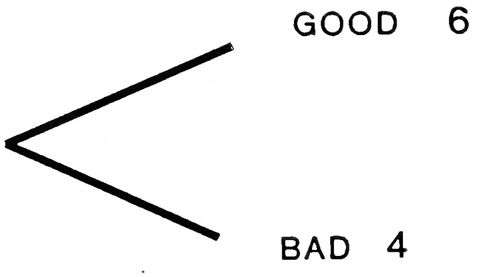


FIGURE 6

DIV* Decision Tree

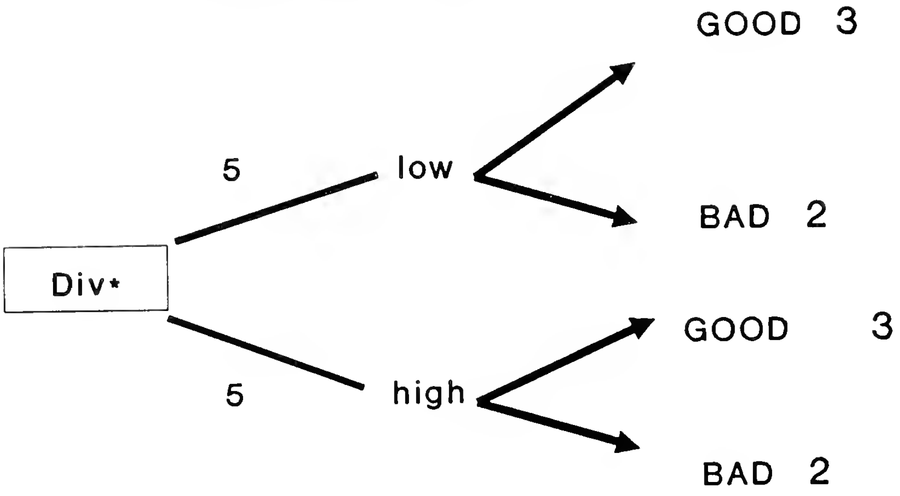


FIGURE 7

NOF* Decision Tree

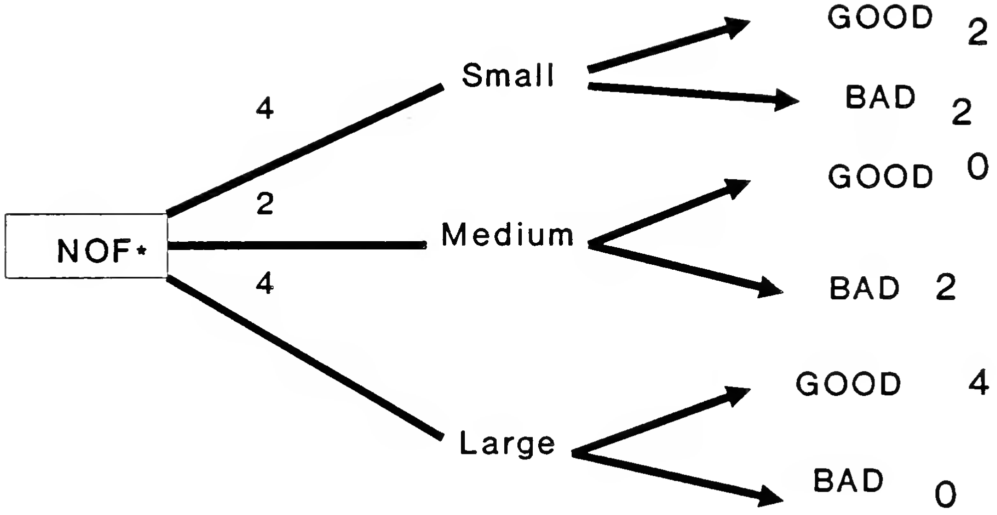


FIGURE 8

NIF* Decision Tree

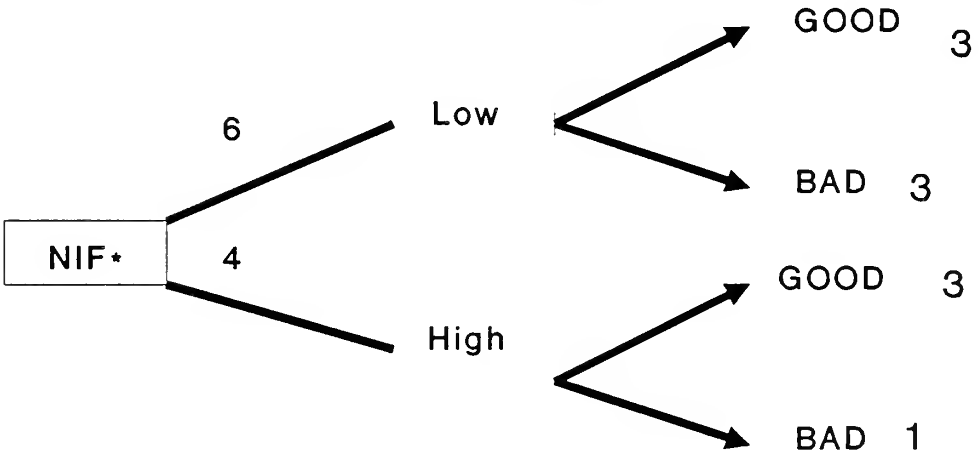


FIGURE 9
NOF* and DIV* Decision Tree

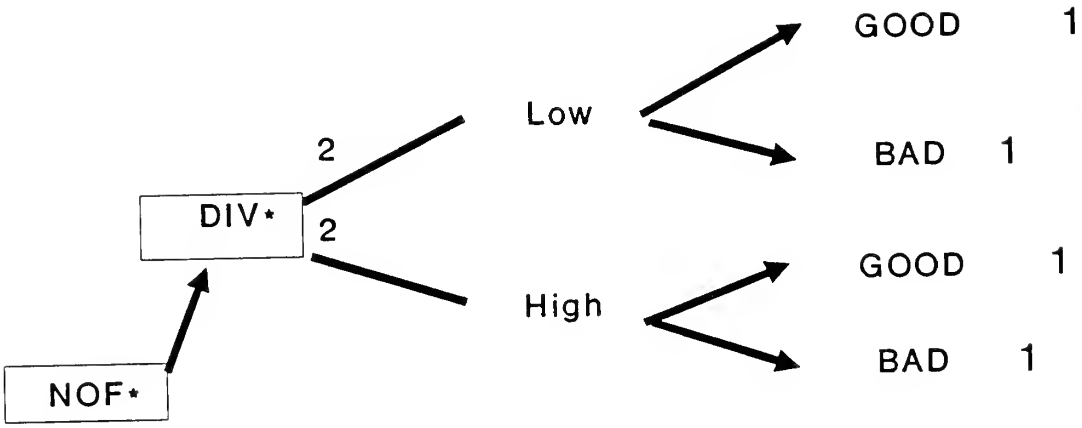


FIGURE 10
NOF* and NIF* Decision Tree

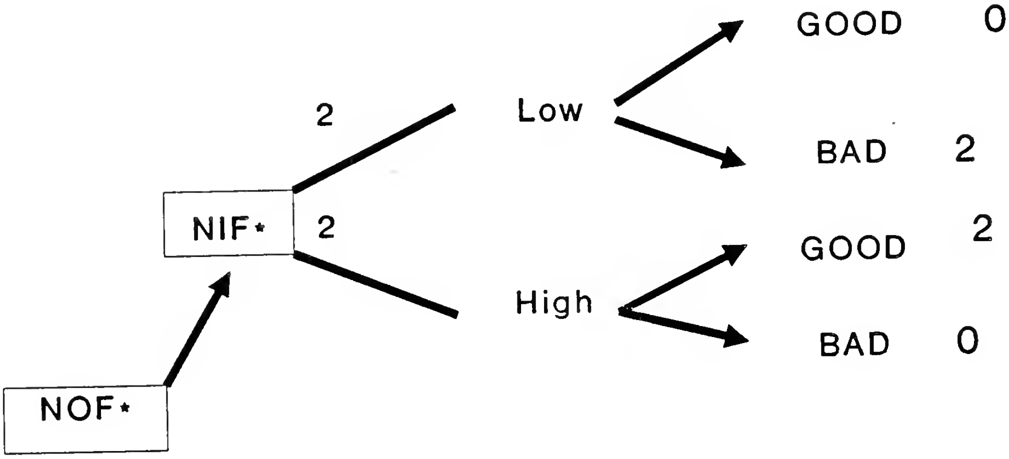
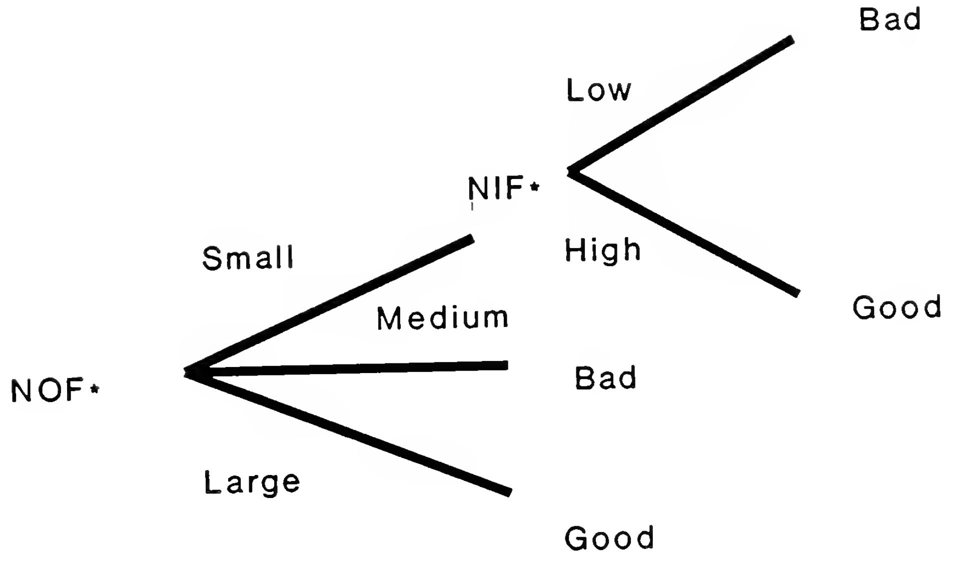


FIGURE 11

Final Decison Tree


Bad
Low
NIF*
Small
High
Good
Medium
NOF*
Bad
Large
Good